A. Proof of Dimension Independence for Output Perturbation (Theorem 1)

First, we prove the following lemma, which bounds the excess loss (empirical risk) due parameter vector \( \theta_{priv} \) compared to \( \hat{\theta} \).

**Lemma 1.** Let \( L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell((\theta, x_i); y_i) \). We have,

\[
\mathbb{E}_b \left[ L(\theta_{priv}) - L(\hat{\theta}) \right] = O \left( \frac{(LR_2)^2 \sqrt{\log(1/\delta)} + \epsilon}{\lambda x} \right).
\]

**Proof.** Now,

\[
L(\theta_{priv}) - L(\hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} (\ell((\theta_{priv}, x_i); y_i) - \ell((\hat{\theta}, x_i); y_i)).
\]

By the Lipschitz property of the loss function \( \ell \), we have

\[
L(\theta_{priv}) - L(\hat{\theta}) \leq \frac{1}{n} \sum_{i=1}^{n} L_i|\theta_{priv} - \hat{\theta}, x_i|.
\]

Notice that, each inner product \( \langle b, x_i \rangle \) is distributed as \( \mathcal{N}(0, \sigma^2 \| x_i \|_2) \), where \( \sigma = \frac{(LR_2)\sqrt{\log(1/\delta)} + \epsilon}{\lambda x} \). Therefore,

\[
\mathbb{E}_b \left[ L(\theta_{priv}) - L(\hat{\theta}) \right] \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_b[|\langle b, x_i \rangle|] \leq \frac{L_\sigma}{n} \sum_{i=1}^{n} \| x_i \|_2 \leq LR_2 \sigma.
\]

Hence Proved. \( \square \)

Now, let \( J(\theta) = \mathbb{E}_{(x, y) \sim \text{Dist}}[\ell((\theta, x); y)] + \frac{\lambda}{2n} \| \theta \|_2^2 \) and \( \hat{J}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell((\theta, x_i); y_i) + \frac{\lambda}{2n} \| \theta \|_2^2 \). Also, let \( \theta^* = \arg \min_{\theta \in \mathbb{R}^p} J(\theta) \) and \( \hat{\theta} = \arg \min_{\theta \in \mathbb{R}^p} \hat{J}(\theta) \). Then, using Lemma 1, we have:

\[
\mathbb{E}_b[J(\theta_{priv}) - \hat{J}(\hat{\theta})] \leq O(LR_2 \sigma) + \mathbb{E}_b \left[ \frac{\lambda \| \theta_{priv} \|_2^2}{2n} \right].
\]

Now, we use the following excess risk theorem by (Shalev-Shwartz et al., 2009).

**Theorem 5** (One sided uniform convergence (Shalev-Shwartz et al., 2009)). Let \( J(\theta) \), \( \hat{J}(\theta), \hat{\theta}, \lambda \) and the loss function \( \ell \) be defined as above. Then, the following holds \( \forall \theta \in \mathbb{R}^p \) (with probability at least \( 1 - \gamma \)):

\[
J(\theta) - J(\theta^*) \leq 2 \left( \hat{J}(\theta) - \hat{J}(\hat{\theta}) \right) + O \left( \frac{(LR_2)^2 \log(1/\gamma)}{\lambda} \right),
\]

where \( L \) is the Lipschitz constant of the loss function \( \ell \), and \( R_2 \) is an upper bound on the \( L_2 \)-norm of the feature vectors in the training data set.

Let \( F(\theta) = \mathbb{E}_{(x, y) \sim \text{Dist}}[\ell((\theta, x); y)] \). From Theorem 5 and (13), we have the following with probability at least \( 2/3 \) over the data generating distribution \( \text{Dist} \):

\[
\mathbb{E}_b[J(\theta_{priv}) - J(\theta^*)] \leq O(LR_2 \sigma) + \mathbb{E}_b \left[ \frac{\lambda \| \theta_{priv} \|_2^2}{2n} \right] + O \left( \frac{(LR_2)^2}{\lambda} \right).
\]

That is,

\[
\mathbb{E}_b[F(\theta_{priv}) - F(\theta^*)] \leq O(LR_2 \sigma + \frac{(LR_2)^2}{\lambda}) + \frac{\lambda}{2n} \| \theta^* \|_2^2.
\]

Theorem now follows by using \( \sigma = \frac{(LR_2)\sqrt{\log(1/\delta)} + \epsilon}{\lambda x} \) in the above given bound and by using Markov’s inequality.

B. Proofs for Private ERM over Simplex

**B.1. Proof of Privacy Guarantee (Theorem 3)**

**Proof.** We first characterize the optimal non-private \( \hat{\theta} \) obtained by solving (8). To this end, we form the Lagrangian of (8):

\[
L(\theta, \nu) = \frac{1}{n} \sum_{i=1}^{n} \ell((\theta, x_i); y_i) + \frac{\lambda}{n} \sum_{j=1}^{p} \theta_j \log(\theta_j) + \nu \left[ \sum_{i=1}^{r} \theta_i - 1 \right]
\]

Now, using optimality conditions:

\[
(\hat{\theta}, \nu^*) = \max_{\nu} \min_{\theta \in \Delta} L(\theta, \nu).
\]

By setting the gradient of the Lagrangian to be zero and using primal feasibility, we get:

\[
\hat{\theta}_j = \exp \left( -\frac{\nu^*}{\lambda} - 1 - \frac{1}{\lambda} \sum_{i=1}^{r} \ell'(\langle x_i, \hat{\theta} \rangle; y_i)x_i^j \right),
\]

\[
\exp \left( \frac{\nu^*}{\lambda} \right) = \sum_{r \in [p]} \exp \left( -1 - \frac{1}{\lambda} \sum_{i=1}^{r} \ell'(\langle x_i, \hat{\theta} \rangle; y_i)x_i^r \right),
\]

where \( \ell' \) is the derivative of \( \ell \) and \( x_i^j \) denotes the \( j \)-th coordinate of \( x_i \).

That is,

\[
\hat{\theta}_j = \frac{\exp \left( -\frac{1}{\lambda} \sum_{i=1}^{r} \ell'(\langle x_i, \hat{\theta} \rangle; y_i)x_i^j \right)}{\sum_{r \in [p]} \exp \left( -\frac{1}{\lambda} \sum_{i=1}^{r} \ell'(\langle x_i, \hat{\theta} \rangle; y_i)x_i^r \right)}.
\]
Similarly, let $\hat{\theta}_j^*$ be the solution to (8) but by using a different data set $D'$ that differs from $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ in exactly one data point. Without loss of generality, we assume that $D$ and $D'$ differs only in the first entry $(x_1', y_1')$. Now, consider an index $a_s$ that is sampled from the probability distribution $\hat{\theta}$. Now, probability of sampling $a_s = j$, given that $\hat{\theta}$ is learned using data set $D$ is given by: $Pr(a_s = j | D) = \hat{\theta}_j$. Similarly, $Pr(a_s = j | D') = \hat{\theta}_j'$. Hence,

$$\max_j Pr(a_s = j | D) = \max_j \hat{\theta}_j$$

$$= \max_j \frac{\exp \left( -\frac{1}{\lambda} \sum \ell'(\langle x_i, \hat{\theta} \rangle; y_i)x_i \right)}{\sum \exp \left( -\frac{1}{\lambda} \sum \ell'(\langle x_i, \hat{\theta} \rangle; y_i)x_i \right)}$$

Adding the above two equations, using the fact that $D - D' = (x_1', y_1')$, by applying the Lipschitz continuity of $\ell$, and by using Holder’s inequality, we get:

$$||\hat{\theta} - \hat{\theta}'||_1 \leq \frac{LR_{\infty}}{\lambda}.$$ 

Now plugging the above bound in (17), we get:

$$A \leq \exp \left( \frac{4LR_{\infty}}{\lambda} + \frac{2nLR_{\infty}^3L_g}{\lambda^2} \right).$$

Using the above equation with (16), we get:

$$\max_j Pr(a_s = j | D') \leq \exp \left( \frac{4LR_{\infty}}{\lambda} + \frac{2nLR_{\infty}^3L_g}{\lambda^2} \right).$$

Note that this ensures, that each “sample” $a_s$ is $\epsilon = \exp \left( \frac{4LR_{\infty}}{\lambda} + \frac{2nLR_{\infty}^3L_g}{\lambda^2} \right)$ differentially private. Hence, $\epsilon$ and $(\epsilon, \delta)$ differential privacy for the computation of the collection of $m$ samples ($a_1, a_2, \ldots, a_m$) and consequently $\theta_{\text{priv}}$ follows by using the weak and the strong composition theorems of (Dwork et al., 2006b; 2010c) respectively.

**B.2. Proof Utility Guarantee (Theorem 4)**

We first prove in Lemma 2 the excess risk bound of Algorithm (8) for any choice of $m$ and $\lambda$. We then set $m = \left( \frac{\epsilon \lambda}{\log(1/\delta)} \right)^2 \left( 32 + \frac{16nR^2 L_g}{\lambda^2} \right)^{-2}$ and $\lambda = \frac{n^{2/3}}{\epsilon^{1/3} \log(1/\delta) \cdot p}$ to get the final guarantee.

**Lemma 2.** Let $L, L_g, \gamma$ be as defined in Theorem 3. With probability at least $2/3$ over the randomness of $\text{Dist}$ and the randomness of $\theta_{\text{priv}}$, the following is true.

$$\mathbb{E}_{(x, y) \sim \text{Dist}} [\ell(\langle \theta_{\text{priv}}, x \rangle; y) - \ell(\langle \theta^*, x \rangle; y)] = O \left( \frac{LR_{\infty} \log m}{\sqrt{m}} + \frac{\lambda}{n} \log p + \frac{(LR_{\infty})^2}{\lambda} \right).$$

Here $\theta^* = \arg \min_{\theta \in \Delta} \mathbb{E}_{(x, y) \sim \text{Dist}} [\ell(\langle \theta, x \rangle; y)]$.

**Proof.** Recall that,

$$\theta_{\text{priv}} = \frac{1}{m} \sum_s e_{a_s},$$

where $e_{a_s}$ is the $a_s$-th canonical basis vector and $a_s \in \{1, 2, \ldots, p\}$, $\forall s \in [m]$ are sampled i.i.d. according to the probability distribution $\hat{\theta}$.

Now, for any fixed $x$: $\langle x, \theta_{\text{priv}} \rangle = \frac{1}{m} \sum_s \langle x, e_{a_s} \rangle$. Note that, $\mathbb{E}_{a_s} [\langle x, e_{a_s} \rangle] = \langle x, \hat{\theta} \rangle$. Therefore,

$$\mathbb{E}_{\theta_{\text{priv}}} [\langle x, \theta_{\text{priv}} \rangle] = \mathbb{E}_{a_s} [\langle x, e_{a_s} \rangle] = \langle x, \hat{\theta} \rangle$$
Furthermore, $|\langle x, e_a \rangle| \leq \|x\|_\infty = R_\infty$. Therefore by Hoeffding’s inequality, with probability at least $1 - \gamma$,

$$|\langle x, \theta_{priv} \rangle - \langle x, \hat{\theta} \rangle| = O \left( \frac{R_\infty \log(1/\gamma)}{\sqrt{m}} \right).$$

Observing $|\langle x, \theta_{priv} \rangle - \langle x, \hat{\theta} \rangle|$ is universally bounded by $R_\infty$, and setting $\gamma = \frac{1}{\sqrt{m}}$, we have

$$E_{\theta_{priv}} \left[ |\langle x, \theta_{priv} \rangle - \langle x, \hat{\theta} \rangle| \right] = O \left( \frac{R_\infty \log m}{\sqrt{m}} \right).$$

Now,

$$E_{\theta_{priv}} \left[ E_{x \sim \text{Dist}} \left[ |\langle x, \theta_{priv} \rangle - \langle x, \hat{\theta} \rangle| \right] \right] =$$

$$\leq \max_{x \in X} E_{\theta_{priv}} \left[ |\langle x, \theta_{priv} \rangle - \langle x, \hat{\theta} \rangle| \right] = O \left( \frac{R_\infty \log m}{\sqrt{m}} \right).$$

Therefore, with probability at least $9/10$ over the randomness of $\theta_{priv}$, we have

$$E_{(x,y) \sim \text{Dist}} \left[ \ell((x, \theta_{priv}); y) - \ell((x, \hat{\theta}); y) \right] = O \left( \frac{LR_\infty \log m}{\sqrt{m}} \right).$$

Now, using standard uniform convergence bound of (Shalev-Shwartz et al., 2009; Kakade et al., 2008), we get:

$$E_{(x,y) \sim \text{Dist}} \left[ \ell((x, \hat{\theta}); y) - \ell((x, \theta^*); y) \right] =$$

$$O \left( \frac{LR_\infty \log m}{\sqrt{m}} + \frac{\lambda n \log p + (LR_\infty)^2}{\lambda} \right).$$

□